VU VRIJE UNIVERSITEIT AMSTERDAM

VU Research Portal

When a taxonomy meets data

Hartman, C.A.; Hox, J.; Mellenbergh, G.J.; Boyle, M.H.; Offord, D.R.; Racine, Y.; McNamee, J.; Gadow, K.D.; Sprafkin, J.; Kelly, K. L.; Nolan, E.E.; Tannock, R.; Schachar, R.; Schut, H.; Postma, I; Drost, R.; Sergeant, J.A.

published in
Journal of Child Psychology and Psychiatry
2001

DOI (link to publisher)
10.1111/1469-7610.00778

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)
Hartman, C. A., Hox, J., Mellenbergh, G. J., Boyle, M. H., Offord, D. R., Racine, Y., McNamee, J., Gadow, K. D., Sprafkin, J., Kelly, K. L., Nolan, E. E., Tannock, R., Schachar, R., Schut, H., Postma, I., Drost, R., & Sergeant, J. A. (2001). When a taxonomy meets data. *Journal of Child Psychology and Psychiatry*, *42*(6), 817-836. https://doi.org/10.1111/1469-7610.00778

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 05. May. 2023

J. Child Psychol. Psychiat. Vol. 42, No. 6, pp. 817–836, 2001
Cambridge University Press
© 2001 Association for Child Psychology and Psychiatry
Printed in Great Britain. All rights reserved
0021–9630/01 $15.00 + 0.00

DSM-IV Internal Construct Validity: When a Taxonomy Meets Data

Catharina A. Hartman
University of Groningen, The Netherlands

Joop Hox
Utrecht University, The Netherlands

Gideon J. Mellenbergh
University of Amsterdam, The Netherlands

Michael H. Boyle, David R. Offord, Yvonne Racine, and Jane McNamee
McMaster University, Hamilton, Canada

Kenneth D. Gadow, Joyce Sprafkin, Kevin L. Kelly, and Edith E. Nolan
State University of New York, Stony Brook, U.S.A.

Rosemary Tannock and Russell Schachar
Hospital for Sick Children, Toronto, Canada

Harry Schut
Regional Institute for Mental Welfare, The Hague, The Netherlands

Ingrid Postma
Regional Institute for Mental Welfare, Hoorn, The Netherlands

Rob Drost
De Jutter, The Hague, The Netherlands

Joseph A. Sergeant
Free University, Amsterdam, The Netherlands

The use of DSM-IV based questionnaires in child psychopathology is on the increase. The internal construct validity of a DSM-IV based model of ADHD, CD, ODD, Generalised Anxiety, and Depression was investigated in 11 samples by confirmatory factor analysis. The factorial structure of these syndrome dimensions was supported by the data. However, the model did not meet absolute standards of good model fit. Two sources of error are discussed in detail: multidimensionality of syndrome scales, and the presence of many symptoms that are diagnostically ambiguous with regard to the targeted syndrome dimension. It is argued that measurement precision may be increased by more careful operationalisation of the symptoms in the questionnaire. Additional approaches towards improved conceptualisation of DSM-IV are briefly discussed. A sharper DSM-IV model may improve the accuracy of inferences based on scale scores and provide more precise research findings with regard to relations with variables external to the taxonomy.

Keywords: Classification, construct validity, DSM, factor analysis, questionnaires, symptomatology.

Abbreviations: CBCL: Child Behavior Checklist; CD: Conduct Disorder; CFA: Confirmatory Factor Analysis; CFI: Comparative Fit Index; CSI-4: Child Symptom Inventory-4; EPC: expected paramenter change; GFI: Goodness of Fit Index; ML: maximum likelihood estimation; OCHS-R: Ontario Child Health Study Scales-Revised; ODD: Oppositional Defiant Disorder; RMR: root mean square residual; RMSEA: root mean square error of approximation; TRF: Teacher Report Form.

Introduction

Conceptualisation of child psychopathological disorders is dominated by a clinically based perspective, such as that formulated in the *Diagnostic and statistical manuals of mental disorders* (DSM-III, DSM-III-R, DSM-IV, American Psychiatric Association, 1980, 1987, 1994). Parallel to this clinical viewpoint is a psychometric perspective on how to conceptualise childhood disorders, which has also enjoyed considerable success (Krueger, Caspi, Moffitt, & Silva, 1998). This has resulted in two different taxonomies of childhood disorders: "clinical syndromes", which originate from hypotheses about

Requests for reprints to: Catharina A. Hartman, University of Groningen, Faculty of Medical Sciences, Dept. of Psychiatry, PO Box 30001, 9700 RB Groningen, The Netherlands (E-mail: c.a.hartman@med.rug.nl).

covarying symptoms derived from observations of patients by clinicians (Wakefield, 1999), and "empirical syndromes", which are empirically generated on the basis of statistical covariation between symptoms without a priori conceptions of what the important constituents of the taxonomy should be (Achenbach, 1991 a). The former taxonomy is generally used in clinical diagnostic interviews, whereas the latter is used in questionnaires. Although the clinical and empirical perspectives share the background assumption on syndromes as co-occurring patterns of symptoms (Achenbach, 1995; Wakefield, 1999), little attempt has been made to address the relative internal construct validities of these two perspectives (Loevinger, 1957; Skinner, 1981). Although we do not question the potential utility of competing models to describe the phenomenology of childhood disorders in order to improve accuracy, it seems that these different taxonomies exist next to one another merely because of past traditions. This is unfortunate, since we regard a common *conceptual* framework for the domain of child psychopathology as a prerequisite for the scientific *understanding* of that domain (Wakefield, 1999).

Several investigators have made a strong case that future improvement of classification in (child) psychiatry must be based on an integration of the clinical and psychometric perspectives (Clark, Watson, & Reynolds, 1995; Kamphaus & Frick, 1996; Waldman, Lilienfield, & Lahey, 1995). The present paper provides one such integrative attempt by investigating the DSM-IV taxonomy, as operationalised in questionnaires, on the basis of psychometric principles derived from factor analysis. As such, the present research cuts directly into two long-standing differences between clinical and empirical taxonomies:

(1) A categorical as opposed to a dimensional model of child psychopathology, where the clinical perspective has been tied to the former and the psychometric to the latter, and
(2) A conceptually derived as opposed to an empirically derived model of child psychopathology, where, again, the clinical perspective has been tied to the former and the psychometric to the latter.

However, these differences need not necessarily hold up the integration of the clinical and psychometric perspectives. The approach taken here, i.e. using a dimensional approach in order to empirically test the internal construct validity of a conceptually derived taxonomy, is not in conflict with these long-standing differences, for the following two reasons.

First, the observed phenomenology of childhood disorders can reasonably be viewed as having dimensional qualities (see, for example, Bannister, 1968; Cantwell & Rutter, 1994; Cromwell, 1975; Tennen, Hall, & Affleck, 1995). This argues against a preconceived adoption of the categorical model. In contrast, preliminary adoption of a dimensional model is not in conflict with the potential existence of categories. Waller and Meehl (1998) argue that it is a misconception that a discrete and qualitatively different category (a "taxon") precludes dimensionality. The convenient dichotomy "taxonic versus dimensional", should read, strictly speaking, "taxonic-dimensional versus dimensional only". That is, when a qualitatively distinct category exists, its distribution is likely to be part of a mixture, located within one or more latent dimensions. Thus, if a set of indicators have appreciable validity, regardless of whether the underlying construct is categorical or dimensional, they covary, and a dimensional approach such as factor analysis must necessarily reveal a factor (Meehl, 1999). In the absence of knowledge on categories that are qualitatively different from normality, the factor analytic model may be used for determining distinguishable problem domains and their valid indicators.

Second, although the psychometric approach toward syndrome conceptualisation resulted from dissatisfaction with the initial lack of empirical validation of the clinical taxonomy (Achenbach, 1995; Achenbach & Edelbrock, 1978; Quay, 1986a, b), the distinction empirical versus nonempirical is no longer meaningful: both taxonomies aim at empirical validation. The current difference between clinical and empirical syndromes is best characterised by the former representing a deductive, i.e. a conceptually driven, yet possibly empirically modified approach, and the latter representing an inductive, data-driven approach towards conceptualisation of childhood psychopathology. A truly inductive analysis is difficult to achieve, due to the many subjective decisions that are required in its use, such as which variables to include in the analysis or how many factors to extract (Block, 1995), which influence the outcome of the analysis considerably. In contrast, reliance on a priori knowledge may result in more appropriate recovery of the factors and their relative positions toward each other than when purely data-driven methods are used, as was recently shown by Little, Lindenberger, and Nesselroade (1999).

The present paper, therefore, focuses on the clinically derived DSM-IV syndromes, measuring the concepts they purport to measure using the dimensional method of factor analysis in a deductive manner. The reason for analysing DSM-IV in the present study is pragmatic: we do not know of any better phenomenological description of child psychopathology. It should be emphasised at this point that internal construct validity is only one aspect of construct validity. The hallmark of construct validity is external construct validity, for example through differential relations of current clinical concepts with aetiology, course, prognosis, or dysregulations in the neurobiological or cognitive systems. The present research is important, since the better the internal construct validity, the greater potential there is to find differential relations with variables external to the taxonomy.

Method

Subjects

Data were collected from the Netherlands, Canada, and the United States of America. Table 1 provides the age and gender distributions for each of these samples. Both parent and teacher ratings of both clinically referred and general population samples were analysed, except for the United States, for which the parent general population sample was missing.

Instruments

The following questionnaires were used: the Ontario Child Health Study Scales-Revised (OCHS-R) for Canada (Boyle et al., 1993; Macleod, McNamee, Boyle, Offord, & Friedrich, 1999), and the Child Symptom Inventory-4 (CSI-4) for the United States (Gadow & Sprafkin, 1994, 1997). These questionnaires were conceptually developed to represent a number of DSM syndrome dimensions. For the Netherlands, the Child Behavior Checklist (CBCL) and the Teacher Report Form (TRF) were used (Achenbach, 1991a, b), along with a

Table 1
Sample Characteristics[a]

	Canada: OCHS-R				Netherlands: CBCL(TRF)/DSM-IV				United States: CSI-4		
	Pop		Clinic		Pop		Clinic		Pop	Clinic	
	Parent	Teacher	Parent	Teacher	Parent	Teacher	Parent	Teacher	Teacher	Parent	Teacher
Age range	6–17	6–17	3–19	3–19	4–13	4–13	3–19	3–17	5–13	3–19	3–19
Boys	876	815	1020	702	456	456	541	432	790	671	637
Girls	879	850	681	396	509	509	264	205	733	235	215
Total	1775	1665	1701	1098	965	965	805	637	1523	906	852

[a] Given by country, instrument used, type of sample (Pop: population sample; Clinic: clinically referred sample), and informant (parent/teacher).

questionnaire that was designed to measure the DSM-IV constructs studied here (see below). Thus, the Dutch DSM-IV model evaluated here consists of items drawn from the CBCL and the TRF, along with additional DSM-IV targeted items.

Since the questionnaires of the present study were independently developed, they differ in a number of respects. Although these differences in operationalisation may account for some of the variation in the results, it should be borne in mind that the focus of the present study is on the validity of the hypothesised DSM concepts. The outcome of a factor analytic study is influenced by the idiosyncrasies of the variables, the sample, or the informant included in the analysis. Strong evidence for the internal construct validity may be inferred when these latent variables are corroborated in multiple operationalisations of the constructs, in multiple samples, and when rated by multiple informants. This cannot be achieved on the basis of a single operationalisation, a single type of sample, or a single type of informant.

These instruments measure multiple constructs. Multidimensional models are more diagnostic than unidimensional models in demonstrating internal construct validity. By contrasting a particular construct with what it is not, e.g. Major Depression is not measured by items that measure Generalised Anxiety, the meaning of the constructs in the model become more circumscribed and, hence, the model can be more easily disconfirmed by data (see Gerbing & Anderson, 1984). We studied 6 constructs measured in 11 samples: Problems with Attention, Hyperactivity-Impulsivity, Conduct Disorder (CD), Oppositional Defiant Disorder (ODD), Generalised Anxiety, and Depression. Our choice to focus on a common latent structure should not be confused with the scope of the questionnaires. These six constructs form a subset of what is covered by the OCHS-R, the CSI-4, and the CBCL/TRF. The OCHS-R additionally includes items measuring Separation Anxiety. The CSI-4 additionally includes items measuring Separation Anxiety Disorder, Obsessive Compulsive Disorder, Tourette syndrome, Schizophrenia, Pervasive Developmental Disorder, Social Phobia, and Bipolar Disorder. The CBCL/TRF consists of items that measure the empirically derived constructs Withdrawn, Somatisation, Anxious/Depressed, Social Problems, Thought Problems, Attention Problems, Delinquency, and Aggression. The Dutch DSM-based items cover the six constructs studied here.

DSM-IV Model

A model is a statement about the associations between variables. The DSM-IV model specifies which items are associated with which syndrome dimensions. Here the DSM-IV model is tested against the data by means of factor analysis. The basic assumption is that DSM-IV syndrome dimensions are latent variables (factors) whose manifestations are behavioural symptoms. These latent variables are assumed to be the organising force underlying the observed response consistencies on the behavioural symptoms in the questionnaire. In other words, the items in a syndrome scale should measure a common factor in order to be valid. The model specifies that the covariance among the manifest items is explained by the factor loadings of items on common factors and by the correlation between these factors. Confirmatory Factor Analysis (CFA), used here, aims at empirical verification of the assumed relations of items with DSM-IV constructs. In contrast to an EFA, both the number of factors and the orientation of the factors in the factorial hyperspace are clearly defined by the model.

The factor loadings are estimated from the data. A factor loading represents the degree to which an item is an indicator of the latent syndrome dimension. This feature allows the items to be imperfect indicators of the underlying construct. It also allows for a probabilistic rather than a defining nature of the items as indicators of an underlying syndrome dimension. That is, the factor analytic model is consistent with the polythetic principle of DSM that guides case definition. The polythetic principle holds that none of the symptoms which are hypothesised to be indicators of the underlying syndrome is either sufficient or necessary for caseness (or a high score, in dimensional terms).

The correlations between the factors are also estimated from the data. This allows for the known positive association between severity of a disorder and the likelihood that a child will meet the criteria for another disorder (Kovacs & Devlin, 1998). It also allows for the possibility that some syndrome dimensions are more related to one another than others (Krueger et al., 1998; Newman et al., 1996).

If the model adequately fits the data, and items have been shown to be adequate indicators of the underlying construct, items may be summed into a scale score that may be interpreted as the degree to which the syndrome is present. This is similar to the prototypical principle of DSM. The prototypical principle holds that patients can be ranked with regard to their degree of category membership or prototypical resemblance (Klein, 1999).

If the model is consistent with the data, the correlations between factors may be interpreted as the degree to which the different problem dimensions co-occur.

Data Analysis

Childhood psychiatric symptoms do not fulfil the factor analytic requirements of normally distributed variables. They are generally skewed (see, for example, Farrington & Loeber, in press). There is no agreed best method for factor analysing a large number of highly skewed, ordinally scored items with sample sizes as used here. For reasons described in Hartman et al. (1999) and Hartman (2000), here, maximum likelihood estimation (ML) was applied to covariances.

Model Fit

Conventional rules of fit. When variables are skewed and categorical rather than distributed normally, the chi-square statistic in ML estimation does not follow the theoretical chi-square distribution but is inflated. This seriously impedes the evaluation of the adequacy of the models studied here. Additional fit indices are considered in the present study: root

Table 2
Model Fit for Comparative Factor Models in the Canadian, Dutch, and U.S. Samples

	Indep. model 1	1-factor model 2	2-factor model 3	3-factor model 4	DSM model 5	99% Interval model 5	DSM+ model 6	Unrestr. model 7
Canada								
Parents								
Clinic ($N = 1701$)								
df	3003	2925	2920	2917	2898	2898	2863	2550
χ^2	63086	36251	25044	21458	18313	2842–3311	14549	10193
RMSEA	.11	.082	.067	.061	.056	.000–.009	.049	.042
RMR	.23	.12	.086	.082	.075	.018–.021	.057	.028
GFI	.19	.40	.60	.67	.73	.95–.96	.79	.85
CFI	.00	.50	.63	.69	.74	.99–1.00	.81	.87
Pop ($N = 1775$)								
df	3003	2925	2920	2917	2898	2898	2863	2550
χ^2	51924	22006	17354	15152	13309	3865–4623	11015	8223
RMSEA	.10	.061	.053	.049	.045	.014–.018	.040	.035
RMR	.23	.069	.060	.057	.053	.021–.024	.043	.027
GFI	.23	.62	.73	.78	.81	.94–.95	.85	.88
CFI	.00	.61	.70	.75	.79	.96–.98	.83	.88
Teachers								
Clinic ($N = 1098$)								
df	2485	2414	2409	2406	2388	2388	2361	2074
χ^2	51731	27886	19883	16423	12978	2507–2925	10507	7802
RMSEA	.13	.10	.081	.073	.064	.007–.014	.056	.050
RMR	.30	.13	.10	.093	.081	.021–.026	.058	.029
GFI	.14	.34	.51	.61	.70	.93–.94	.76	.81
CFI	.00	.48	.65	.72	.78	.99–1.00	.83	.88
Pop ($N = 1665$)								
df	2485	2414	2409	2406	2388	2388	2361	2074
χ^2	77217	35980	27533	23163	18356	4355–5461	14898	11221
RMSEA	.13	.091	.079	.072	.063	.022–.028	.056	.051
RMR	.32	.11	.091	.085	.074	.023–.028	.053	.029
GFI	.12	.40	.54	.63	.73	.91–.93	.78	.82
CFI	.00	.55	.66	.72	.79	.95–.97	.83	.88
Netherlands								
Parents								
Clinic ($N = 805$)								
df	5050	4949	4944	4941	4923	4923	4885	4459
χ^2	42457	25478	19774	17163	14609	5292–5970	12716	10588
RMSEA	.10	.072	.061	.055	.049	.010–.016	.045	.041
RMR	.22	.11	.081	.075	.071	.028–.032	.056	.033
GFI	.16	.38	.54	.62	.68	.88–.89	.73	.77
CFI	.00	.45	.60	.67	.64	.96–.99	.79	.84

Pop ($N = 965$)								
df	5050	4949	4944	4941	4923	4923	4885	4459
χ^2	43364	22687	19840	18116	16356	8051–10019	14697	12670
RMSEA	.089	.066	.056	.053	.049	.026–.033	.046	.044
RMR	.21	.087	.064	.060	.058	.034–.039	.050	.035
GFI	.18	.50	.64	.69	.73	.83–.86	.76	.79
CFI	.00	.33	.61	.66	.70	.84–.89	.74	.79
Teachers								
Clinic ($N = 637$)								
df	4465	4370	4365	4362	4345	4345	4300	3910
χ^2	42933	25155	20329	17315	14584	5177–5808	12304	10019
RMSEA	.12	.086	.076	.068	.061	.017–.023	.054	.050
RMR	.28	.12	.10	.092	.087	.031–.035	.066	.034
GFI	.12	.31	.42	.51	.61	.84–.86	.67	.72
CFI	.00	.46	.59	.66	.73	.95–.97	.79	.84
Pop ($N = 965$)								
df	4371	4277	4272	4269	4252	4252	4208	3822
χ^2	50948	30867	25571	22836	19686	8911–11508	16680	14214
RMSEA	.11	.080	.072	.067	.061	.034–.042	.055	.053
RMR	.23	.10	.086	.080	.075	.037–.044	.062	.038
GFI	.16	.40	.51	.58	.66	.79–.84	.71	.74
CFI	.00	.43	.54	.60	.67	.80–.87	.73	.78
United States								
Parents								
Clinic ($N = 906$)								
df	1711	1652	1647	1643	1626	1626	1608	1327
χ^2	28345	17813	15413	11978	7932	1709–2091	6506	5056
RMSEA	.13	.10	.10	.083	.065	.008–.018	.058	.054
RMR	.24	.12	.12	.10	.077	.025–.030	.059	.033
GFI	.23	.41	.47	.55	.72	.93–.94	.77	.82
CFI	.00	.39	.48	.61	.76	.98–1.00	.82	.86
Teachers								
Clinic ($N = 852$)								
df	1540	1484	1479	1476	1459	1459	1442	1219
χ^2	32307	18522	16379	12438	7737	1569–1952	6545	4267
RMSEA	.15	.12	.11	.093	.071	.009–.020	.064	.054
RMR	.30	.13	.13	.11	.080	.024–.029	.061	.023
GFI	.17	.36	.40	.52	.70	.92–.94	.73	.86
CFI	.00	.45	.52	.64	.80	.98–1.00	.83	.92
Pop ($N = 1523$)								
df	1081	1034	1029	1027	1011	1011	1004	814
χ^2	55369	27801	24144	17152	7966	2150–3262	6970	4974
RMSEA	.18	.13	.12	.10	.067	.027–.038	.062	.058
RMR	.36	.12	.11	.10	.065	.023–.320	.050	.023
GFI	.14	.35	.39	.50	.79	.91–.94	.82	.86
CFI	.00	.51	.57	.70	.87	.95–.98	.89	.92

χ^2 is rounded to the nearest integer; *df*: degrees of freedom; Indep.: Independence model; DSM+: DSM-IV modified model; Unrestr.: Unrestricted model; Clinic: clinically referred sample; Pop: population-based sample.

mean square error of approximation (RMSEA) (Steiger, 1990), root mean square residual (RMR) (Bollen, 1989), Goodness of Fit Index (GFI) (Jöreskog & Sörbom, 1989; Tanaka & Huba, 1985), and the Comparative Fit Index (CFI) (Bentler, 1990). These fit indices are based on different construction principles and consequently emphasise different aspects of model fit (Hu & Bentler, 1999). Rules of thumb are generally used for the range of values which are taken to indicate adequate fit. The ranges are: RMSEA (0.03–0.08); RMR (0–0.05); GFI (0.90–1.00); CFI (0.90–1.00). Whether these rules of thumb apply to the present situation (large models, large sample sizes, and categorically skewed variables, resulting in a less than optimal measure of association—covariance—and estimation method —ML) is unknown. Therefore, the chi-square and the fit indices are interpreted in the present study with the aid of two unequivocal means of assessing model fit: simulation and comparison with other models. These are described below.

Simulation. A simulation study provides empirical distributions for the various fit indices taking into account the skewed, categorical distributions of the item responses, as observed in the samples. In addition, potential effects of both model size and sample size are incorporated in these distributions. To accomplish this, simulation samples with the distribution characteristics observed in the data are repeatedly drawn from a population for which the model being evaluated holds, but with the introduction of random error through sampling. The model under study is fitted to each of these simulation samples, in order to obtain an empirical sampling distribution of the fit indices. Actual values of the fit indices as they are found for the DSM-IV model in each of the samples used here may then be compared with this range of values which fall under random sampling variations when the model is consistent with the data, given model size and sample size. In these simulated distributions of the fit indices, potentially inadequate fit due to inaccuracy of the model is disentangled from apparent inadequate fit caused by distributional violations. Thus, these empirical sampling distributions of the fit indices provide a test of the fit of the DSM-IV model.

In summary, the simulation study was designed here such that (1) model size and sample size for the simulation samples are identical to model and sample size in the actual data for the DSM-IV model; (2) the distribution characteristics of the items in the simulation samples are like the item responses in the sample for the DSM-IV model; and (3) the simulation samples are drawn from a population for which the covariance structure implied by the DSM-IV model holds.

To obtain precise results (Efron & Tibshirani, 1993), 400 simulation samples were drawn for each of the 11 samples. For each sample, an empirical probability distribution is provided for chi-square, RMSEA, RMR, GFI, and CFI, based on 400 fits of the DSM-IV model to these simulation samples (see Boomsma, 1983; Hartman et al., 1999; Hox & Hartman, 1999, for further details). The Simulcat (Hox, 1998) and EQS 5.6 programs (Bentler, 1995) were used.

Comparison with other models. A second means of judging adequacy of model fit is to compare the DSM-IV model with other models. Four models were considered in which fewer problem dimensions were posited than the six dimensions in the DSM-IV model. This allows for an evaluation of the explanatory power of the DSM-IV model above and beyond these five more "crude" representations of the structure in the data. In addition, two models were considered which are less restrictive than the DSM-IV model. Comparison with the goodness of fit of these models evaluates the degree to which the DSM-IV model failed to represent accurately the covariance structure in the data.

Model 1, the most restrictive model fitted to the data, is the independence model. Model 1 hypothesises that all items in the model are uncorrelated, indicating that no common factors underlie the items. The independence model has the lowest possible fit compared to models that do assume common factors. It is thus the baseline for evaluating fit of other models.

Model 2 is the single-factor model. This model tests the possibility that a single undifferentiated latent dimension describes the covariance structure of the items.

Model 3 is a two-factor model that distinguishes between internalising and externalising problems (Achenbach & Edelbrock, 1978; Cantwell, 1996; Rutter et al., 1969; Verhulst & Van der Ende, 1992). The first factor consists of the items measuring Problems with Attention, Hyperactivity-Impulsivity, ODD, and CD. The second factor consists of the items measuring Generalised Anxiety and Depression. These two factors were allowed to correlate (Angold, Costello, & Erkanli, 1999; Wångby, Bergman, & Magnusson, 1999).

Model 4 is a three-factor model in which ADHD, Aggressive Behaviour, and Internalising Problems are separate factors (Achenbach, 1991a, b; Wångby et al., 1999) and allowed to correlate.

Model 5 is the DSM-IV model. The six factors were allowed to correlate. The degree to which the DSM-IV model shows improved fit over and above the goodness of fit of the aforementioned models provides insight into its relative explanatory power to describe the covariance patterns in the data.

Model 6 is a modified DSM-IV model. When the DSM-IV model did not provide good fit in each of these samples, it was explored post hoc where the factor loading matrix deviated. Modification was based on the expected parameter change indices (EPCs). The EPC index indicates those items that load with a factor at a higher level than is predicted from the loadings of items on common factors and the correlation between these factors, which are estimated on the basis of the DSM-IV model. Thus, the EPC indices were used to identify the necessary but not a priori specified loadings in the DSM-IV model (see Kaplan, 1990). These loading were modelled, starting with the largest, and up to a loading of 0.20. Clearly, the DSM-IV modified model which resulted from this data-driven procedure capitalises on chance. Despite the fact that part of the secondary factor loadings reported in the Results section (Tables 3a to 3f) ought to be due to chance fluctuations, this procedure was considered the best possible option for evaluating how well the DSM-IV model held up once the items were free to load with any factor that improved consistency with the covariance structure.

Model 7, the unrestricted model (Jöreskog, 1979), is the least restricted model fitted to the data. In the unrestricted model, no specific pattern is specified for the items loading with the underlying syndrome dimensions, except for the minimum number of restrictions required for model identification (Jöreskog, 1979). Model 7 assesses whether the number of factors describe the data adequately, regardless of the content of the scales. The fit of the unrestricted model indicates the best possible fit for a six-factor model.

All models were fitted to the data using Lisrel 8.12a (Jöreskog & Sörbom, 1993).

Results

Part 1: Model Fit

Aptness of the DSM-IV model: Simulation. Table 2, column 7 provides the results of the simulation study. For each fit index a 99% two-sided interval was derived. Each interval is based on 400 fits to 400 simulation samples. The intervals in Table 2, column 7 encompass the range of values that indicate good fit given the specific properties of the samples. They are used to evaluate goodness of fit of the DSM-IV model estimated from the data (Table 2, column 6).

All fit indices based on the DSM-IV model were outside the 99% range in all samples. These results indicate that the DSM-IV model did not fit the covariance structure here.

Aptness of the DSM-model: Comparison with alternative models. The DSM-IV model was compared with six

models: the independence model, the single-factor, a two-factor, a three-factor, a DSM-modified model, and the unrestricted model. Table 2 provides goodness of fit for these alternative models in columns 2, 3, 4, 5, 8, and 9, respectively.

The independence model provides, by definition, the lowest possible fit, since it tests the hypothesis that there are no common factors underlying the responses to the items. The results showed a poor fit throughout (Model 1). In each sample there was substantial covariance among the items, which may be explained by common factors.

The single-factor model showed considerable improvement in fit compared with the independence model (Model 2). This indicates that a substantial part of the covariation is explained by a single undifferentiated factor. This finding was salient in the population-based Canadian and Dutch samples, and more for the parent population samples than the teacher population samples. The clinical Canadian and Dutch samples were less influenced by this undifferentiated factor. The single-factor model accounted for substantially less covariance in the three U.S. samples, compared with the Canadian and Dutch samples.

The two-factor model (externalising and internalising) showed substantial enhancement in model fit compared with the single-factor model in all samples (Model 3). For the U.S. samples, model fit improvement of the two-factor model compared to the single-factor model was poorer compared with the Canadian and Dutch samples. For the Canadian and Dutch samples, model fit improvement was more pronounced for the clinical than for population samples. The implication is that, with the specification of two factors, more covariance is explained for the Dutch and Canadian samples than for the U.S. samples.

The three-factor model (Externalising, Internalising, and ADHD as separate factors) showed considerable improvement compared to the two-factor model for all samples (Model 4). In contrast to the single- and two-factor models, improvement in model fit was largest for the three U.S. samples.

The DSM-IV model (Model 5) distinguished six factors: Problems with Attention, Hyperactivity-Impulsivity, CD, ODD, Generalised Anxiety, and Depression. The DSM-IV model showed clear enhancement in fit compared with the three-factor model for all samples. This finding was somewhat more pronounced for teacher than for parent data. Improvement in fit was greatest for all U.S. samples, in particular, for teachers in the population sample.

The DSM-IV modified model (Model 6) showed a consistent improvement in fit (RMR). This model showed a substantial number of secondary loadings above and beyond the primary loadings present in the DSM-IV model. Since the DSM-IV modified model was the result of a data-driven optimalisation procedure conducted on each sample separately, enhancement in model fit was similar across samples. Relative to the loss in degrees of freedom, the U.S. samples gained most from this modification.

The unrestricted model (Model 7) demonstrated that six factors were not adequate to explain the covariance in the data. The unrestricted model showed about an equally good fit for the Canadian and US samples, but somewhat poorer fit for the Dutch samples. Parent data showed better fit than teacher data for both Canada and The Netherlands. In contrast, the U.S. teacher data showed better fit than the parent data.

In summary, the hypothesised DSM-IV model was corroborated by a consistent increase in model fit with the specification of additional factors in all samples. The covariance structure in the US samples was most consistent with the DSM-IV model compared with the Canadian and Dutch samples: there was a relative greater improvement in fit for the U.S. samples over and above that of the internalising and externalising problem domains. This finding indicates that the evidence that ADHD is separate from internalising and externalising problems, and that Problems with Attention, Hyperactivity-Impulsivity, CD, ODD, Generalised Anxiety, and Depression are separate from one another, was most pronounced in the U.S. samples. Two sources were identified to explain why the DSM-IV model accounted for the covariance in the data in neither of the samples. First, the DSM modified model showed that model fit benefited from the specification of a substantial number of secondary factor loadings in all countries. Second, the unrestricted model showed that the six factors were not adequate to fully account for the covariance here.

Part II: Substantive Results

Factor loadings. Tables 3a to 3f provide the factor loadings of the DSM-IV modified model for Problems with Attention, Hyperactivity-Impulsivity, CD, ODD, Generalised Anxiety, and Depression, respectively.

The first column lists all items. Within each problem domain, those items that were measured in all 11 samples are provided first, followed by items common to The Netherlands and Canada, and those common to The Netherlands and the United States. Finally, items are provided measured in a single country.

The remaining columns of Tables 3a to 3f are structured as follows: columns 2 to 6 describe the Canadian samples, 7 to 11 the Dutch samples, and 12 to 15 the U.S. samples.

Within each country, the first four columns (three for the U.S.) provide the degree to which each item measures a given problem domain for the clinical parent data, the population parent data (not available for the United States), the clinical teacher data, and the population teacher data. Thus, four columns of factor loadings are provided for the Canadian, four for the Dutch, and three for the U.S. samples.

Occasionally, the number of factor loadings is smaller, e.g., when items were measured in the parent samples, but not in the teacher samples (e.g. "trouble sleeping"). Within each country, the last column provides the problem dimensions for which secondary loadings were observed (6th, 11th, and 15th columns), aggregated across the four (three for the U.S.) samples.

There are two types of secondary loadings: those specified a priori, indicated by a * (see legends in Tables 3a and 3f), and those modelled in the second instance, through post hoc model modification. Only those problem domains are listed in Tables 3a to 3f for which the secondary loadings were $\geqslant 0.20$. Where the frequency of cross-loadings exceeds the number of samples, the implication is that in one or more samples more than one secondary loading was present.

The final column (column 16) in Tables 3a to 3f lists the average factor loadings based on all samples where an item was measured, as an index of each of the item's

Table 3a
Factor Loadings for the DSM-IV Modified Model in the Canadian, Dutch, and U.S. Samples: Problems with Attention

	Canada					Netherlands					United States				Total
Problems with Attention	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	tc[c]	tp[d]	Secondary loadings	Mean
Can't pay attention for long*	.78	.80	.88	.88		.80	.82	.81	.83		.76	.83	.89		.83
Does not seem to listen	.43	.46	.79	.83	(2 odd)[a,b]	.15	.21	.43	.28	(2 h-i[a,d]; 2 odd[a,b];1 cd[c])	.50	.51	.66	(3 h-i)[a,c,d]	.48
Fails to finish things	.65	.65	.74	.76		.71	.61	.74	.71		.81	.82	.89		.74
Loses things	.40	.37	.60	.67	(1 cd[a];1 odd[b])	.43	.44	.62	.38	(1 h-i[d]; 1 cd[a];1 dep[b])	.68	.63	.81		.55
Distractible/trouble sticking to anything	.79	.57	.88	.84	(1 h-i)[b]	.86	.83	.88	.86						.81
Difficulty following directions/instructions	.69	.71	.77	.77		.42	.40	.54	.48	(4 odd)[a,b,c,d]					.60
Jumps from one activity to another*	.47	.51	.56	.55	(3 h-i)[a,c,d]	.25	.24	.39	.30	(4 h-i)[a,b,c,d]		.43		(1 h-i)[c]	.41
No attention to details/careless mistakes						.55	.62	.61	.61		.72	.68	.80		.66
Difficulty organising work and activities						.54	.39	.70	.68	(1 dep)[b]	.75	.83	.89		.68
Avoids tasks which require mental effort						.62	.63	.68	.65		.66	.70	.82		.68
Easily distracted by other things going on						.80	.79	.65	.63	(2 h-i)[c,d]	.71	.59	.86	(1 h-i)[c]	.72
Is forgetful in daily activities						.54	.43	.54	.55	(2 dep)[b,c]	.76	.71	.85		.63
Inattentive, easily distracted						.87	.83	.86	.82						.85

pc[a] = parent, clinically referred sample; pp[b] = parent, population sample; tc[c] = teacher, clinically referred sample; tp[d] = teacher, population sample.
Secondary loadings: odd: ODD; cd: CD; h-i: Hyperactivity/Impulsivity; dep: Depression; anx: Generalised Anxiety; att: Problems with Attention.
* Specified a priori to load additionally with another factor: "Can't pay attention for long" with Generalised Anxiety and Depression; "Jumps from one activity to another" with Hyperactivity/Impulsivity.

Table 3b
Factor Loadings for the DSM-IV Modified Model in the Canadian, Dutch, and U.S. Samples: Hyperactivity/Impulsivity

	Canada					Netherlands					United States				Total
Hyperactivity/Impulsivity	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	tc[c]	tp[d]	Secondary loadings	Mean
Can't stay seated when required	.80	.72	.56	.80	(1 att)[c]	.69	.54	.77	.69		.78	.79	.68	(1 att)[d]	.71
Restless/jittery/hyperactive*	.79	.78	.56	.81	(1 att)[c]	.74	.64	.58	.51	(3 att)[b,c,d]	.70	.73	.73	(1 anx)[a]	.69
Has difficulty playing quietly	.68	.63	.78	.80		.44	.34	.73	.32	(3 att[a,b,d]; 1 cd[d];1 dep[c])	.77	.79	.85		.65
Talks excessively	.51	.54	.68	.70		.60	.61	.78	.69		.64	.73	.79		.66
Interrupts/blurts out answers	.48	.65	.80	.80	(1 odd)[a]	.58	.52	.74	.69		.56	.73	.83		.67
Difficulty awaiting turn in games	.69	.66	.80	.81		.70	.70	.80	.75		.68	.84	.87		.75
Interrupts or butts in on others	.39	.43	.82	.81	(2 odd)[a,b]	.73	.45	.57	.43	(1 att[c]; 2 cd[c,d];1 odd[b])	.59	.84	.85	(1 odd)[a]	.63
Fidgets	.40	.41	.33	.38	(4 att)[a,b,c,d]	.66	.67	.59	.65	(1 att[c]; 1 dep[a])					.51
Impulsive/acts without thinking	.11	.34	.62	.81	(2 att[a,b]; 3 odd[a,b,c])	.49	.37	.45	.55	(3 att[a,b,d]; 1 cd[c])	.16			(1 att; 1 odd)[a]	.43
Jumps from one activity to another*	.30	.19	.26	.27	(4 att)[a,b,c,d]	.48	.39	.40	.36	(4 att)[a,b,c,d]		.36		(1 att[c])	.33
Does dangerous things*	.23	.29	.30	.28	(4 cd)[a,b,c,d]	.49	.38	.22	.21	(4 cd)[a,b,c,d]	.25	.45		(1 cd[c]; 1 odd[a])	.31
Moves with hands or feet/squirms in seat						.68	.67	.59	.42	(2 att)[c,d]	.70	.54	.53	(2 att)[c,d]	.59
Is "on the go" or acts if "driven by a motor"						.75	.67	.78	.73		.76	.75	.84		.75
Constant chatting in class or during meals						.63	.57	.74	.42	(1 att)[d]					.59
Talks out of turn						.72	.70	.84	.76						.76
Disrupts class discipline						.72	.69	.83	.81						.76
Gets increasingly restless during day						.52	.50	.65	.42	(1 att)[d]					.52
Runs about or climbs on things						.68	.44	.52	.43	(3 cd)[b,c,d]					.52
Needs a lot of supervision												.43		(1 att; 1 odd)[c]	.43

For abbreviations see Table 3a.

* Specified a priori to load additionally with another factor: "Restless/jittery/hyperactive" with Generalised Anxiety and Depression; "Jumps from one activity to another" with Problems with Attention; "Does dangerous things" with CD.

Table 3c
Factor Loadings for the DSM-IV Modified Model in the Canadian, Dutch, and U.S. Samples: Conduct Disorder

	Canada					Netherlands					United States				Total
Conduct Disoder	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	tc[c]	tp[d]	Secondary loadings	Mean
Cruel/bullying/mean	.24	.29	.58	.38	(4 odd)[a,b,c,d]	.22	.31	.79	.72	(2 odd)[a,b]	.36	.47	.53	(3 odd)[a,c,d]	.44
Gets in many fights	.35	.58	.81	.78	(1 odd)[a]	.25	.44	.78	.61	(1 h-i; 1 odd)[a]	.41	.65	.61	(3 odd)[a,c,d]	.57
Uses weapons when fighting	.50	.37	.49	.55		.33	−.07	.42	.03	(2 dep)[b,d]	.45	.44	.34		.35
Physically attacks people	.58	.53	.85	.81		.26	.43	.76	.68	(2 odd)[a,b]	.30	.89	.81	(1 odd)[a]	.63
Destroys things belonging to others	.54	.58	.69	.66	(1 h-i)[a]	.60	.64	.72	.61	(1 h-i)[a]	.44	.40	.75	(2 odd)[a,c]	.60
Lying or cheating	.50	.39	.43	.47	(4 odd)[a,b,c,d]	.43	.20	.61	.65	(2 odd)[a,b]	.38	.08	.27	(3 odd)[a,c,d]	.40
Steals outside the home	.60	.52	.50	.53		.52	.36	.40	.38		.63	.39	.53		.49
Truancy/plays hookey from school	.20	.36	.03	.07	(3 dep)[a,c,d]	.10	.10	.13	.07	(2 dep)[a,c]	.55	.01	.22	(1 dep)[c]	.17
Does dangerous things*	.49	.33	.41	.47	(4 h-i)[a,b,c,d]	.32	.24	.55	.41	(4 h-i)[a,b,c,d]	.16	.30		(2 h-i[a,b]; 1 odd[a])	.37
Cruel to animals	.34	.32	.21	.38		.34	.28	.26	.14		.34				.29
Threatens people	.38	.64	.83	.81	(1 odd)[a]	.58	.41	.66	.52						.60
Sets fires	.46	.33	.24	.32		.48	.30	.25	.31		.34				.34
Vandalism	.64	.45	.62	.56		.63	.61	.64	.60						.59
Hangs around kids in trouble	.56	.57	.59	.51	(1 att)[a]	.42	.20	.50	.35	(1 anx)[b]					.46
Destroys own things	.51	.58	.48	.61	(1 att[c]; 1 h-i[a])	.48	.63	.56	.55	(1 h-i)[a]					.55
Swearing or obscene language	.35	.30	.72	.50	(3 odd)[a,b,d]	.28	.28	.72	.50	(3 odd)[a,b,d]		.18		(1 odd)[c]	.43
Broken into house/building/car	.41	.30				.23	.02	.22			.60				.30
Steals things at home	.63	.47				.62	.21	−.01	−.03	(2 odd[c,d]; 1 anx[d]; 1 dep[b])					.32
Runs away from home	.28	.31			(1 dep)[a]	.39	.07				.61				.33
Stolen things using physical force						.14	.05	.31	.38		.30	.49	.65		.33
Stays out at night when not supposed to						.30	.06				.61				.32
Forced someone into sexual activity											.29				.29
Engaged in illegal/unlawful activities											.50				.50
Irresponsible school/work/money											.07			(1 att; 1 odd)[a]	.07
Does not seem to care about suffering others											.21			(1 odd)[a]	.21
Tries to actually hurt others in a fight												.91			.91
Breaks important rules appropriate for age												.28		(1 odd)[c]	.28

For abbreviations see Table 3a.
* Specified a priori to load additionally with another factor: "Does dangerous things" with Hyperactivity/Impulsivity.

Table 3d
Factor Loadings for the DSM-IV Modified Model in the Canadian, Dutch, and U.S. Samples: Oppositional Defiant Disorder

	Canada					Netherlands					United States				Total
Oppositional Defiant Disorder	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	tc[c]	tp[d]	Secondary loadings	Mean
Temper tantrums or hot temper	.75	.52	.82	.80	(1 dep)[b]	.70	.59	.47	.65	(1 cd)[c]	.72	.82	.82		.70
Argues a lot with adults	.71	.69	.82	.83		.74	.51	.40	.34	(2 h-i[b,d]; 2 cd[c,d])	.75	.81	.79		.67
Does things that annoy others	.57	.55	.24	.40	(1 att[a]; 3 h-i[b,c,d]; 1 cd[c])	.55	.45	.11	.10	(4 cd)[a,b,c,d]	.72	.77	.81		.48
Blames others for own mistakes	.65	.45	.55	.51	(1 att[b]; 2 h-i[c,d])	.66	.63	.30	.44	(1 h-i[c]; 2 cd[c,d])	.71	.72	.81		.58
Easily annoyed by others	.65	.39	.75	.55	(1 h-i[d]; 1 dep[b])	.47	.39	.77	.66	(2 dep)[a,b]	.64	.75	.78	(1 dep)[a]	.62
Angry/resentful*	.60	.41	.64	.65	(4 dep)[a,b,c,d]	.62	.44	.83	.76	(2 dep)[a,b]	.69	.77	.81	(2 dep)[a,c]	.66
Gets back at people	.40	.25	.21	.27	(4 cd)[a,b,c,d]	.30	.11	.28	.34	(4 cd[a,b,c,d]; 1 dep[b])	.78	.67	.84	(1 cd)[c]	.40
Cranky*	.46	.44	.56	.56	(4 dep)[a,b,c,d]	.36	.22	.67	.64	(3 dep)[a,b,c]	.30	.54	.36	(3 anx)[a,c,d]	.46
Disobedient at school	.11	.14	.52	.47	(4 h-i[a,b,c,d]; 2 cd[a,b])	.11	.17	.22	.08	(3 h-i[a,b,d]; 4 cd[a,b,c,d])					.23
Defiant/talks back to adults	.77	.75	.85	.83		.80	.67	.38	.33	(1 h-i[d]; 1 cd[c])					.67
Explosive and unpredictable	.73	.23	.81	.55	(2 cd[b,d]; 1 dep[b])	.74	.38	.37	.27	(2 h-i[b,d]; 2 cd[c,d])					.51
Defies or refuses						.78	.75	.51	.57	(2 cd)[c,d]	.74	.78	.82		.71
Acts stubborn						.64	.65	.74	.73			.76			.70
Persistent testing of limits						.56	.50	.13	.00	(4 h-i[a,b,c,d]; 2 cd[c,d])					.30
Easily wronged						.51	.45	.78	.74	(2 dep)[a,b]					.62
Demands must be met immediately						.69	.66	.60	.73	(1 h-i)[c]					.67
Breaks minor rules												.56		(1 h-i)[c]	.56

For abbreviations see Table 3a.
* Specified a priori to load additionally with another factor. "Angry/resentful" with Depression; "Cranky" with Generalised Anxiety and Depression.

Table 3e
Factor Loadings for the DSM-IV Modified Model in the Canadian, Dutch, and U.S. Samples: Generalised Anxiety

	Canada					Netherlands					United States				Total
Generalised Anxiety	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	tc[c]	tp[d]	Secondary loadings	Mean
Worries about doing better	.68	.63	.73	.65		.66	.71	.73	.68		.25	.50	.60	(1 dep)[a]	.62
Difficulty controlling worries	.77	.51	.60	.60	(3 dep)[b,c,d]	.52	.43	.75	.38	(3 dep)[a,b,d]	.38	.75	.77	(1 dep)[a]	.57
Nervous/high-strung/tense	.25	.13	.31	.39	(3 h-i[a,c,d]; 4 dep[a,b,c,d])	.44	.37	.53	.43	(2 h-i[a,c]; 2 dep[b,d])	.82	.66	.81	(1 h-i)[c]	.47
Overtired*	.10	.01	−.10	−.09	(4 dep)[a,b,c,d]	.15	.16	.10	.08	(2 dep)[a,c]	−.09	−.35	−.21	(3 dep)[a,c,d]	−.02
Cranky*	.02	.03	−.02	.01	(4 odd; 4 dep)[a,b,c,d]	−.05	.02	−.05	−.04	(4 odd[a,b,c,d]; 3 dep[a,b,c])	.46	.28	.43	(3 odd)[a,c,d]	.10
Can't pay attention for long*	.00	−.09	.04	−.02	(4 att)[a,b,c,d]	−.05	−.04	−.03	−.02	(4 att)[a,b,c,d]	.09	.09	.02	(3 att)[a,c,d]	.00
Restless/jittery/hyperactive*	.00	−.05	.01	.01	(1 att[c]; 4 h-i[a,b,c,d])	−.01	−.04	.02	.03	(3 att[b,c,d]; 4 h-i[a,b,c,d])	.20	−.05	.01	(3 h-i)[a,c,d]	.01
Needs constant reassurance	.47	.22	.52	.38	(4 att[a,b,c,d]; 1 dep[b])	.49	.20	.52	.26	(1 h-i[a]; 2 dep[b,d])					.38
Worries about past behaviour	.58	.27	.41	.40	(3 dep)[b,c,d]	.38	.17	.38	.17	(2 dep)[b,d]					.35
Worries about doing wrong	.70	.66	.73	.73		.39	.54	.36	.58	(2 h-i)[a,c]					.59
Is afraid of making mistakes	.75	.71	.74	.73		.62	.72	.60	.74						.70
Too fearful or anxious	.46	.30	.47	.49	(4 dep)[a,b,c,d]	.51	.29	.67	.48	(2 dep)[b,d]					.46
Worries about future	.64	.55	.47	.55	(1 dep)[c]	.70	.54	.73	.34	(1 dep)[d]					.57
Feels he/she has to be perfect	.61	.55	.61	.52		.42	.49	.38	.36						.49
Trouble sleeping*	.13	.06			(2 dep)[a,b]	.22	.20			(1 dep)[b]	.49				.22
Worries a lot about health	.36	.39	.29	.18	(3 dep)[a,c,d]										.31
Is overly anxious to please people	.54	.57	.62	.64											.59
Avoids school to stay home*	.02	−.05	−.03	−.14	(4 dep[a,b,c,d]; 1 cd[b])										−.05
Is anxious or worried several times a day						.59	.10	.68	.29	(2 dep)[b,d]					.42
Afraid of new things or situations						.56	.57	.57	.62						.58
Shy or timid						.40	.30	.23	.43	(1 dep)[c]					.34
Acts restless or edgy											.63	.52	.62	(3 h-i)[a,c,d]	.59

For abbreviations see Table 3a.

* Specified a priori to load additionally with another factor: "Overtired", "Trouble sleeping", and "Avoids school to stay home" with Depression; "Cranky" with ODD and Depression; "Can't pay attention for long" with Problems with Attention and Depression; "Restless/jittery/hyperactive" with Hyperactivity/Impulsivity and Depression.

Table 3f
Factor Loadings for the DSM-IV Modified Model in the Canadian, Dutch, and U.S. Samples: Depression

	Canada					Netherlands					United States				Total
Depression	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	pp[b]	tc[c]	tp[d]	Secondary loadings	pc[a]	tc[c]	tp[d]	Secondary loadings	Mean
Unhappy/sad/depressed	.55	.62	.74	.72	(1 anx)[a]	.76	.76	.30	.76	(1 anx)[c]	.70	.76	.72		.67
Little enjoyment pleasurable activities	.62	.46	.65	.63		.70	.52	.48	.52	(1 anx)[c]	.62	.64	.65		.59
Feels worthless or inferior	.41	.61	.56	.60	(3 anx)[a,c,d]	.34	.45	.03	.47	(1 odd[c]; 4 anx[a,b,c,d])	.75	.67	.69		.51
Talks about killing self	.48	.38	.40	.20		.37	.35	.00	.25	(1 anx)[c]	.39	.35	.37	(1 cd)[a]	.32
Self-conscious/easily embarrassed	.19	.14	.23	.21	(4 anx)[a,b,c,d]	.10	.02	−.02	.03	(4 anx)[a,b,c,d]	.77	.68	.70		.28
Feels hopeless	.54	.66	.74	.72	(1 anx)[a]	.35	.48	.11	.59	(1 odd[c]; 3 anx[a,b,c])	.81	.66	.78	(1 odd)[c]	.59
Overtired*	.47	.40	.55	.56		.33	.19	.37	.19		.67	.82	.77		.49
Cranky*	.26	.22	.30	.28	(4 odd)[a,b,c,d]	.50	.39	.24	.14	(4 odd)[a,b,c,d]	.18	.12	.13	(3 odd; 3 anx)[a,c,d]	.25
Can't pay attention for long*	−.06	.02	−.07	−.05	(4 att)[a,b,c,d]	.03	−.04	.04	.03	(4 att)[a,b,c,d]	−.15	−.15	−.05	(3 att)[a,c,d]	−.04
Restless/jittery/hyperactive*	−.02	−.01	−.09	.00	(1 att[c]; 4 h-i[a,b,c,d])	−.08	−.12	−.11	−.10	(3 att[b,c,d]; 4 h-i[a,b,c,d])	−.02	.03	.04	(3 h-i[a,c,d]; 1 anx[a])	−.04
Angry/resentful*	.28	.43	.31	.28	(4 odd)[a,b,c,d]	.26	.24	−.03	.03	(4 odd)[a,b,c,d]	.23	.20	.16	(3 odd)[a,c,d]	.22
Underactive/low energy	.49	.41	.49	.55		.50	.41	.64	.31	(1 att)[d]					.48
Feels too guilty	.11	.31	.24	.27	(4 anx)[a,b,c,d]	.05	.37	−.11	.15	(4 anx)[a,b,c,d]					.17
Has difficulty making mistakes	.14	.01	.23	.13	(4 att; 4 anx)[a,b,c,d]	.13	.23	.12	.08	(1 att[a]; 3 anx[b,c,d])					.13
Complains of loneliness	.23	.48	.54	.45	(1 anx)[a]	.28	.53	.07	.55	(2 anx)[a,c]					.39
Feels that no one loves him	.19	.45	.59	.51	(2 odd[a,b]; 1 anx[a])	.21	.58	.00	.48	(3 odd[a,c,d]; 2 anx[a,c])					.38
Deliberately harms self/suicide	.36	.22	.38	.15		.09	.14	.14	.17	(1 cd)[a]					.21
Lost a lot of weight	.20	.17				.16	.18								.18
Gained a lot of weight	.24	.23				.17	.24								.22
Sleeps more than most kids	.42	.25				.19	.13								.25
Trouble sleeping*	.32	.33				.18	.30			(2 anx)[a,b]	.02			(1 anx)[a]	.23
No interest in usual activities	.58	.46	.44	.47	(2 att)[c,d]										.49
Has trouble enjoying self	.69	.61	.73	.74											.69
Avoids school to stay home*	.46	.24	.38	.49	(1 cd)[b]										.39
It is hard to cheer him/her up						.65	.53	.57	.37	(2 odd)[c,d]					.53
Sorrowful						.73	.69	.42	.76	(1 anx)[c]					.65
Languid						.60	.48	.84	.52						.61
Listless, doesn't feel like doing anything						.68	.48	.79	.43						.60
Withdrawn, does not get involved with others						.21	.42	.59	.50	(1 anx)[a]					.43

For abbreviations see Table 3a.

* Specified a priori to load additionally with another factor: "Overtired", "Trouble sleeping", and "Avoids school to stay home" with Generalised Anxiety; "Cranky" with ODD and Generalised Anxiety; "Can't pay attention for long" with Problems with Attention and Generalised Anxiety; "Restless/jittery/hyperactive" with Hyperactivity/Impulsivity and Generalised Anxiety; "Anxiety/resentful" with ODD.

overall performance as an indicator of the targeted construct.

First, the results are described with regard to the comprehensiveness of the scales, i.e. the degree to which all hypothesised facets of a syndrome dimension were empirically corroborated by the data. Second, results are described with regard to the specificity of the scales.

Factor loadings: Comprehensiveness of the scales. Substantial factor loadings for almost all items in the scales were found for the syndrome dimensions Problems with Attention, Hyperactivity-Impulsivity, and ODD. This result suggests that these syndrome dimensions are consistent with the DSM-IV model. In contrast, for CD, Generalised Anxiety, and Depression, a subset of the a priori hypothesised items had low factor loadings.

For CD, it should be noted that a number of items had low variance in the present samples. For example, this was the case for "uses weapons when fighting", "cruel to animals", or "sets fires", which had also low loadings throughout. Conceptually, these items are highly indicative of CD, yet the present results do not support this conclusion. It is, therefore, likely that a number of CD items could not be evaluated on the basis of the present data due to low variance. Based on the present data, and judged by the magnitude of the factor loadings, the CD problem dimension encompasses the continuum fighting–destroying–stealing.

For Generalised Anxiety and Depression, factor loadings were not restricted by low variance of the items. Therefore, the present results indicated that, empirically, Generalised Anxiety and Depression are more narrowly defined than was hypothesised. Items which typically had low loadings on Generalised Anxiety across all samples were "overtired", "cranky", "can't pay attention for long", "restless/jittery/hyperactive", "trouble sleeping", and "avoids school to stay home". These items were hypothesised a priori to load with multiple factors, and were shown here to contribute little to the Generalised Anxiety construct. Likewise, the items "cranky", "angry/resentful", "can't pay attention for long", and "restless/jittery/hyperactive", hypothesised a priori to be factorially complex, contributed little to the Depression construct. Additionally, items that measure weight or sleep problems tended to have low factor loadings in all samples. Finally, "hesitant/difficulty making decisions" and "feels too guilty" typically loaded with Generalised Anxiety, rather than Depression. For Canada and The Netherlands, the content of the Generalised Anxiety scale was dominated by worries that pertain to failing, in addition to worries about the future and new situations, while the content of the Depression scale was dominated by items that pertain to sadness, reduced pleasure, and reduced energy. In comparison to the Canadian and Dutch results, the U.S. Generalised Anxiety and Depression scales were more consistent with DSM conceptualisations, albeit that the Anxiety scale was a rather short scale to begin with and thus consisted of only a few items with substantial factor loadings.

A final observation with regard to the magnitude of the factor loadings is that, for all samples, the factor loadings of the two internalising dimensions tended to be somewhat lower as compared to the factor loadings of the externalising problem dimensions.

Factor loadings: Specificity of the scales. With respect to the specificity of the scales, i.e. the degree to which the constructs could be differentiated from one another, items tended to have secondary loadings above and beyond their loadings with the targeted problem domain. Thus, a substantial number of items were not specific indicators, which indicates that construct differentiation is not optimal. This finding was somewhat more pointed in Dutch samples than in the Canadian samples. Relatively few secondary loadings were found in the U.S. samples.

Table 4
Mean Correlations between Syndromes across 11 Samples

			DSM-IV		DSM-IV+
			Sum Score	Latent	Latent
H-I	x	Att	.69	.69	.62
CD	x	Att	.46	.44	.35
CD	x	H-I	.59	.59	.49
ODD	x	Att	.52	.52	.43
ODD	x	H-I	.66	.70	.63
ODD	x	CD	.74	.80	.60
Anx	x	Att	.44	.23	.16
Anx	x	H-I	.41	.23	.12
Anx	x	CD	.32	.18	.05
Anx	x	ODD	.46	.30	.23
Dep	x	Att	.52	.37	.34
Dep	x	H-I	.40	.18	.10
Dep	x	CD	.46	.33	.27
Dep	x	ODD	.60	.43	.39
Dep	x	Anx	.78	.73	.54

DSM-IV+: DSM-IV modified model; H-I: Hyperactivity/Impulsivity; Att: Problems with Attention; CD: Conduct Disorder; ODD: Oppositional Defiant Disorder; Anx: Generalised Anxiety; Dep: Depression.

We will consider the typical pattern of cross-loadings for each syndrome dimensions separately. First, with regard to Problems with Attention, the largest number were with Hyperactivity-Impulsivity, followed by ODD. Occasionally, secondary loadings were found with CD or Depression.

Second, for Hyperactivity-Impulsivity, the majority of secondary loadings were with Problems with Attention. Additionally, a substantial number of secondary loadings were with CD or ODD.

Third, a substantial number of CD items additionally measured ODD. To a lesser extent, CD items additionally tapped Hyperactivity-Impulsivity.

Fourth, for ODD, secondary loadings were with CD, Hyperactivity-Impulsivity, and Depression. The relative distribution of cross-loadings across these three domains differed for each country.

Fifth, a large number of secondary factor loadings were present for the Anxiety items. They were almost exclusively with Depression. Where secondary loadings were present with other problem domains, most were hypothesised a priori (e.g. the item "cranky" on ODD) in the DSM-IV model.

Finally, for Depression, the vast majority of secondary loadings concerned Generalised Anxiety. Of the remaining secondary loadings, which were with Problems with Attention, Hyperactivity-Impulsivity, and ODD, the majority were hypothesised a priori in the DSM-IV model.

Overall, the CD and ODD items as well as the Generalised Anxiety and Depression items tended to have the largest number of reciprocal cross-loadings.

Thus, delineation of CD and ODD, and of Generalised Anxiety and Depression, proved the most difficult.

Correlations between syndrome dimensions. Table 4 provides the correlations between the syndrome dimensions, averaged over the 11 samples. The first column specifies all pairwise combinations of the six problem domains. The second column provides the average correlations between *unweighted scale scores* on the basis of the a priori defined DSM-IV model. The third column provides the average correlations between the *latent factors* estimated on the basis of the DSM-IV model. The fourth column provides the average correlations between the *latent factors* based on the DSM-IV modified model.

The pattern of correlation between the syndromes based on a priori scale scores, and that based on a priori DSM latent factors was .94. The pattern of correlations based on the unweighted scale scores and the latent factors estimated by the DSM-IV modified model was .93. Finally, the pattern of correlations based on the latent factors estimated by the a priori DSM-IV model and the DSM-IV modified model was .98. These correlations indicate that partialling out construct-irrelevant variance does not so much affect the rank order of the degree to which problem dimensions are correlated. Rather, these three ways of calculating syndrome intercorrelations affect the magnitude with which the syndrome dimensions are correlated.

The results in Table 4 indicate that, overall, the tendency was that correlations between unweighted scale scores were highest (column 2), followed by those based on the latent factors of the a priori DSM-IV model (column 3), followed by those based on the latent factors of the modified DSM-IV model (column 4).

Unweighted scale scores are based on both construct-specific variance and construct-irrelevant variance. Thus, estimates based on unweighted scale scores are likely to be biased (column 2). When variance that was unique to each of the items was partialled out in the DSM-IV model, the correlation between the factors tended to decrease (column 3). To the extent that the DSM-IV model did not fit the data, construct irrelevant variance was still present in the scores on the factors, which biased the correlations between the factors. It was shown in the previous section that a substantial number of items had a nonspecific relation with the hypothesised factor, which disturbed model fit. Consistent with this, the results showed that when the additional loadings were incorporated in the DSM-IV modified model, the estimates of the correlations between factors decreased (column 4). Given that the nonspecific relations of items to factors were modelled in the DSM-IV modified model, the latter correlations may be considered as the most precise estimates of the correlations between the syndrome dimensions.

The results in Table 4, column 4 indicate that Problems with Attention and Hyperactivity-Impulsivity; ODD and Hyperactivity-Impulsivity; and CD and ODD were most strongly correlated. Somewhat less correlated were Generalised Anxiety with Depression, Hyperactivity-Impulsivity with CD, and Problems with Attention with CD, in decreasing order of magnitude. Typically, the correlations between problem dimensions from the externalising domain and those from the internalising domain were considerably lower. Particularly low were the correlations between Generalised Anxiety and CD, and between Depression and Hyperactivity-Impulsivity. Since these results were based on 11 samples, as well as on an optimalised factor loading matrix, they may be considered as fairly stable lower bound estimates of the correlations between syndromes measured by the questionnaires here.

Discussion

The present paper provides a thorough investigation into the internal construct validity of 6 DSM-IV constructs using CFA, on the basis of 3 different operationalisations and 11 samples. This was one of the ideals described for the DSM-IV field study, which could not be accomplished given the constraints in time and resources (Waldman et al., 1995). The present study encompasses both internalising and externalising problem domains, is based on multiple operationalisations of the syndrome dimensions, evaluates the functioning of each item separately, and includes both parent and teacher samples, as well as population-based and clinically referred samples.

In order to appreciate the results reported here, four points are noted on how DSM-based questionnaires differ from a number of related but fundamental principles of the DSM-IV. Questionnaires, even with sound psychometric properties, are not intended to assign a psychiatric diagnosis. First, they are not intended to replace the deliberated diagnostic decisions of the clinician. Second, DSM criteria are designed to diagnose only those conditions where the symptoms are due to an internal dysfunction of some kind, and not due to a normal response to contextual factors (Wakefield, 1999). Given this goal, questionnaires that address the mere presence of psychopathological symptoms are likely to include false positives, as may be inferred from findings such as the larger prevalence estimates arrived at by questionnaires compared with diagnostic interviews (Swanson et al., 1998). Third, the questionnaires evaluated here are not intended to measure the large number of fine-grained diagnostic categories within the more broadly defined diagnostic syndromes of DSM-IV. For example, onset, duration, or course are not measured, which are variables that define many of the diagnostic categories within the overarching, more broadly defined diagnostic domains in the DSM-IV (Wakefield, 1999; Zuckerman, 1999, p. 44). Fourth, no a priori assumptions are made with regard to discrete boundaries between normality and psychopathology, nor are a priori hierarchical rules for diagnostic priority of one disorder over the other followed. Although the necessary and sufficient rules for case definition, as operationalised by explicit inclusion and exclusion rules in DSM-IV, may be used on the basis of questionnaire scores, case definition as a dichotomous decision is not a primary aim when using questionnaire scores.

The questionnaires used here are characterised as being DSM-based because they share the DSM phenomenological descriptions. The DSM-IV model evaluated here provides the basis for forming scales: the item scores are often added to provide scale scores. Typically, total scores on these scales serve to answer substantive research questions. In the clinic, total scale scores are often used in the form of a profile (see, for example, Gadow & Sprafkin, 1997). Scale scores that deviate from some standard of normality direct the clinician to the main problem areas of a particular child. Instead of assigning a diagnosis, the aim of the questionnaires used here is to provide scale

scores that represent the relative likelihoods that problems from separate problem domains are present, given that parent and teacher informants provide important information on children's emotional and/or behavioural functioning (Achenbach, 1995). In the light of these applications, to what extent can these DSM-questionnaires be said to measure the constructs Problems with Attention, Hyperactivity-Impulsivity, CD, ODD, Generalised Anxiety, and Depression?

The latent structure of these six syndrome dimensions was confirmed through a consistent substantial improvement in model fit with the specification of increasingly refined syndrome dimensions in all samples. This finding suggests that it is to some extent meaningful to sum the items of each of these six constructs in scales. For the U.S. samples, the improvement in model fit over and above the distinction between internalising and externalising was found to be the most substantial, compared with the Canadian and Dutch samples. However, none of the samples provided an adequate fit for the DSM-IV model. Modification of the DSM-IV model, in order to uncover model misfit, confirmed the conclusion that the hypothesised latent structure is essentially correct: when items were allowed to load with any factor that improved consistency with the covariance structure in the data, by and large, items still loaded on the constructs originally designated.

Two primary sources were identified to explain why the DSM-IV models did not meet standards of adequate fit. The first was the fact that many of the items purported to measure the six problem domains had a factorially complex structure. These items had loadings on another factor in addition to that on their original factor and are thus not specific indicators of the construct they are purported to measure. The presence of secondary loadings was more frequent in the Dutch than in the Canadian samples. U.S. samples had substantially fewer secondary loadings.

The second source of inadequate fit emerged from the unrestricted model. In all samples, the fit indices for the unrestricted model suggested that more covariance was present than could be explained by the six factors. This implies that the six constructs are currently not unidimensional.

Multidimensionality of the scales, combined with a substantial number of factorially complex items, indicates that measurement precision of current DSM-based questionnaires is limited. In order to explain this conclusion, as well as to point out directions from which improved measurement precision may come, the two sources of inadequate fit, multidimensionality of the scales and factorially complex items, are discussed below.

Unidimensionality

The implication of multidimensional scales is that apparently identical scores may have different meanings, since they reflect two or more latent variables in some unknown mix. Differences between scores across individuals and within individuals (across time) are then ambiguous. How multidimensionality of scales has occurred is best explained by distinguishing between inadequate *conceptualisation* and inadequate *operationalisation* of constructs. Inadequate conceptualisation refers to the possibility that the model fit is inappropriate because one or more of the six syndrome dimensions in these questionnaires is inherently multidimensional rather than unidimensional. Rethinking conceptualisation requires answers to questions such as: "Are Hyperactivity and Impulsivity expressions of the same underlying problem or should they be considered as fundamentally different, and hence multidimensional?". We did not attempt to answer questions such as these here, and only a small part of this discussion addresses reconceptualisation (described below). This is because (1) our understanding of the nature of underlying dysfunction is still primitive, and (2) a factor analytic finding of multidimensionality is easily caused by small but systematic errors in the operationalisation of the construct which, for the purpose of the present paper, has to be addressed first. Two examples of how errors in operationalisation may have caused the present finding of additional factors beyond the hypothesised DSM-IV constructs are provided here.

First, multidimensionality may be found when a single aspect of the more broadly operationalised syndrome dimension is highly represented in the items of a scale. This creates covariation among those items that are similar in meaning above and beyond the factor they are presumed to measure. In CFA, such errors in operationalisation may be accounted for by post hoc specification of covariation among the unique components of two items in the model. Requiring unique covariance means that additional factors, albeit small ones, are present in the data, which may be added to the model. However, the resulting improved model fit does not provide a solution when using unweighted scale scores. Scales that are disproportionally affected by a single facet of the dimension may cause error in conclusions drawn from the scale scores. One solution is removal of redundant items in a subsequent version of the questionnaire.

Second, errors in operationalisation may be more fundamental. Using Hyperactivity and Impulsivity as an example, multidimensionality may be found because Impulsivity is typically operationalised in a group situation, whereas Hyperacitivity is usually not. Thus operationalised, these items may tap into something above and beyond the strict behavioural characteristics intended for the Hyperactivity-Impulsivity problem domain. In this situation, i.e. when operationalisation error affects simultaneously multiple items, the amount of covarying uniquenesses increases substantially, but increasingly obscures the real issue that the scale measures something other than was intended. This can only be solved by rewriting the items of the construct.

Factorially Complex Items

An additional way to investigate a model's adequacy is on the basis of the substantive parameters of the model, i.e. the magnitude of the factor loadings and the correlations between factors, respectively. For the a priori specified DSM-IV model, except for those items with very low variance, and for those describing physical symptoms, factor loadings were invariably substantial (these factor loadings for the DSM-IV model were not reported). This suggests that most items are good indicators of the construct. In contrast, a number of correlations between problem dimensions tended to be rather high, which suggests low construct differentiation (Table 4, column 3).

However, these conclusions are premature since model fit was inadequate. The pitfall in interpreting factor

loadings and factor correlations as evidence of adequate measurement, in the absence of adequate model fit, is that high loadings and high factor correlations may stem from construct *irrelevant* covariance. When an item is incorrectly assumed to measure a problem domain specifically, the estimated factor loading may be reasonably high due to the presence of high factor correlations. Conversely, interpretation of factor correlations in the absence of adequate model fit may erroneously suggest that factors cannot be differentiated from one another, whereas, in fact, these high factor correlations stem from too few or incorrect factor loadings in the model. Prior to interpretation of the substantive parameters, it was therefore required that the factor loading matrix of the DSM-IV model was more consistent with the data. Empirical model modification of the DSM-IV model showed that the reliable, systematic variance of many items contained both construct specific and construct irrelevant variance. Given that the modified factor loading matrix (Tables 3a to 3f) was more consistent with the data (Table 2, column 8 compared with column 6), the magnitude of the correlations between syndrome dimensions for this modified model decreased accordingly (Table 4, column 4).

The implication of this finding is that a substantial number of items are insufficiently refined for optimal measurement of separate DSM-IV syndrome dimensions. This finding is likely to be a prevailing problem in many questionnaires (and possibly also respondent-based, standardised interviews) of child psychopathology. The interpretation of scale scores based on current instruments suggests more differentiation between problem domains than is actually supported by the data, since these scores include variance that is not specific to the targeted constructs. The implications of factorially complex items are not clearly recognised in the field of child psychopathology. Two examples are noted here.

First, Angold et al. (1999) discuss the finding of high comorbidity in child psychopathology as being, in part, an artefact, due to the symptoms that are shared by different diagnoses. In addition to these literally overlapping items, factorially complex items contribute to the high estimates of comorbidity, at least where these estimates are based on covariances between syndrome scales (see, for example, Hinden, Compas, Howell, & Achenbach, 1997), as was shown in the present paper (Table 4, column 2 compared with column 4).

A second example of the invalidating consequences of factorially complex items is a reduced potential to find evidence of concurrent and discriminant validity (Campbell & Fiske, 1959) in the context of informant agreement. Correlations between different informants' ratings of the same problem behaviour tend to be low, i.e. low concurrent validity (Achenbach, McConaughy, & Howell, 1987). Moreover, correlations between different informant's ratings of *different* problem dimensions, which should be low in order to have evidence of discriminant validity, tend to be hardly any lower than the concurrent validity estimates (Shalev, Hartman, Stavsky, & Sergeant, 1995). This finding of low discriminant validity suggests a lack of differentiation between problem dimensions, due to the presence of construct irrelevant variance, through either assignment of items to the wrong factor or factorially complex items. Low discriminant validity suggests that the usual explanations of situational or informant-specific influences for low concurrent validity do not tell the whole story; i.e. in the situation of inadequate differentiation between problem dimensions due to the presence of factorially complex items, estimates of concurrent and discriminant validity are likely to equalise.

Particularly within the internalising and the externalising domains, child psychopathological syndromes tend to be substantially correlated. Little et al. (1999) illustrated how, in this situation, items of different factors populate adjacent, or even overlapping, regions of the factorial hyperspace, with factorially complex items as a consequence. Specific, albeit conceptually more restricted, measures may be used for improved specificity of scale scores (see, for example, Eley & Stevenson, 1999). However, this could imply ignoring key manifestations of several disorders, with the consequence of sacrificing construct validity. With regard to the literally overlapping symptoms, Angold et al. (1999) emphasise that the real problem is the paucity of research on the *differential* characteristics of the symptoms shared by different disorders. Increased sharpening of symptom content towards its specific manifestation within a given disorder may to some extent be achieved for the factorially complex symptoms as well. Additionally, with regard to questionnaires, the pertinence of precise and unambiguous item wording when using questionnaires should be emphasised. A clinician, on being informed that a child tends to miss classes, will try to clarify whether this is due to rule violation, anxiety, depression, or other reasons. In contrast, questionnaire scale scores are blind to the interpretation of the items by the informant. It was found here that the U.S. samples showed more construct differentiation than other samples. This may, at least in part, indicate a combined influence of (1) the more precise and contextualised description of the targeted problem in each item for the CSI-4 compared to the relatively shorter items of the OCHS-R and the Dutch items, which leads to relatively unequivocal interpretation of items by raters (Block, 1995; Goldberg, 1999; Sandoval, 1981), and (2) the CSI-4 being based on strict DSM-IV criteria as compared with a more broad sampling of the problem domains in Canada and the Netherlands. The strict DSM-IV criteria may, in fact, be the best indicators of the constructs evaluated here. Thus, from this finding it may be inferred that more precise measurement is possible by more precise operationalisation of the DSM-IV constructs.

Reconceptualisation of DSM-IV?

The problem of inadequate model fit found here has been attributed to operationalisation error (e.g. multidimensionality through overrepresentation of a single facet of a construct, factorially complex items through imprecision of item wording). In addition to improved operationalisation, more fundamental factors are likely to play a role in achieving improved model fit, which requires reconsideration of the DSM-IV model itself. Three issues concerning how the DSM-IV model disturbs model fit are noted.

First, with regard to symptom overlap between different DSM-IV syndromes, Angold et al. (1999) noted the difficulty of writing symptoms consistent with their specific manifestation for a particular disorder, since "we may not know much about the specifics of the nonspecific symptoms". Even if we would know the subtle differences in symptoms of different disorders, given the crude

phenomenological descriptions in questionnaires, these may not be easily captured in words, or understood by respondents. Moreover, it is likely that symptomatology overlaps, even if the syndromes are qualitatively different. Pending the identification of syndrome-specific markers of underlying dysfunctions, factorially complex symptoms remain diagnostically ambiguous (Klein, 1999).

To the extent that the removal of diagnostically ambiguous symptoms would lessen the construct validity of a syndrome, their nonspecific nature should be taken account of rather than be defined away. This could be done statistically, for example, as here through the incorporation of double loadings in a model. As argued earlier, this provides more precise estimates of the factors' (co)variance. This results in more precise estimates of the relations with variables external to the taxonomy. When using raw scale scores for the selection of (multiple) phenotypically similar groups, the presence of symptoms that are (equally) reflective of multiple syndrome dimensions in a questionnaire argues for their inclusion in the relevant scales, and subsequent consideration of the relative standing of children on the profile of scores for all syndrome dimensions involved. A more structural approach to account for symptom overlap requires reconceptualisation by models that explicitly account for the common and unique features of separate syndrome dimensions. An example is provided by the tripartite model of Anxiety and Depression (Clark & Watson, 1991; Joiner, Catanzaro, & Laurent, 1996; Watson et al., 1995). This model posits three constructs: Somatic Tension and Arousal, specific to Anxiety; Anhedonia and Low Positive Affect, specific to Depression; and General Distress, which is largely nonspecific. Reconceptualisation potentially results in an improved understanding of *how* common and specific dimensions of DSM-IV constructs relate to external variables (Weiss, Suesser, & Catron, 1998).

A second issue when applying the factor analytic model to the six constructs evaluated here is that operationalisation of Problems with Attention, Hyperactivity-Impulsivity, and ODD is typically more consistent with the psychometric principle of "domain sampling" (Cattell, 1952) than the operationalisation of CD, Generalised Anxiety, or Depression. The principle of domain sampling holds that the indicators of a construct are sampled from a broad universe of possible indicators of the domain. This assumes that all selected indicators are equally potent measures of the construct and that there is only a single factor involved in the scale (Zuckerman, 1999, p. 49). Since the indicators of Problems with Attention, Hyperactivity-Impulsivity, and ODD tend to be more homogeneous, representing a more tightly focused problem domain, they are more consistent with the principle of domain sampling. Representation of the latter constructs as separate factors, both in terms of specificity of the items and unidimensionality, is more consistent with the data. When the principle of domain sampling does not apply, improved model fit may be achieved through reconceptualisation of the constructs. Examples of efforts in this direction are provided by Frick et al. (1993) or Vitiello and Stoff (1997). In these studies, CD was represented as a multidimensional construct.

Third, lack of symptom specificity may partly be due to asymmetric relations between groups of symptoms from different disorders. For example, the differentiation between CD and ODD, and between Depression and Generalised Anxiety, may be due to ODD symptoms being nested in CD, and Generalised Anxiety symptoms being nested in Depression. That is, children with CD often manifest symptoms of ODD as well, but not the reverse (DSM-IV; American Psychiatric Association, 1994, p. 94). Likewise, it has been suggested that Generalised Anxiety symptoms form a common component of Depression, and are more likely to precede Depression than to follow. Thus, an Anxiety Disorder without Depression is common, while a Mood Disorder without Anxiety is rare (Chorpita & Barlow, 1998). Similarly asymmetric relations may exist for hyperactivity symptoms being nested in CD: the presence of hyperactivity symptoms increases the risk for later conduct disturbance, but not the reverse (Rutter, 1996); or for Depression being nested in CD; disruptive behaviour in childhood may lead to additional depressive symptomatology in adolescence, but not the reverse (Loeber & Keenan, 1994; Seeley, Lewinsohn, & Rohde, 1997). Correlation coefficients are not sensitive to the asymmetry of relations (Block, 1995). In the factor model, patterns of asymmetry between certain symptoms from different problem dimensions introduce covariance between these subsets of symptoms, which cannot be accounted for by the correlations between the factors. This implies a decreased coherence of symptoms clustering, which manifests itself through a certain amount of nonspecificity of the indicators to their respective factors.

Boundaries between different syndromes may be sharpened by studying the manifestation of psychopathology in homogeneous age groups, ideally in longitudinal designs. This could tap more directly into the issue of nestedness and how it relates to differential onset and course. As a result, more specific models that account for developmental level can be developed.

In conclusion, the results of the present paper indicated that the DSM-IV model was consistent with the structure of the covariance patterns, as indicated by the improvement in model fit compared with simpler models. However, since the DSM-IV model did not meet the absolute standard of adequate model fit, there is substantial room for improvement. On the basis of current syndrome constructs, measurement precision may be enhanced by greater scrutiny at the operationalisation level, both with regard to the unidimensionality of the scales and the specificity of the items for their respective syndrome dimensions. An improved DSM taxonomy may also come from sharper models which take account of common and specific components of different syndrome dimensions, the multidimensional nature of the underlying construct, and the developmental sensitivity of indicators. The boundary conditions of measurement precision are constrained by current limited knowledge of fundamental distinctions in child psychopathology. Enhanced knowledge in this respect is unlikely to come from improved models of symptom associations alone. Rather, appropriate delineation of distinct syndromes may improve with a progressive understanding of the processes underlying manifest symptomatology (Klein, 1999), at multiple functional levels (Wakefield, 1999), such as the neurobiological or cognitive levels of explanation (Nigg, 2000). Sharper measurement of what we *do* know phenomenologically, as proposed here, may enhance such progress and should be pursued. Internal construct validity of current syndrome conceptualisations remains an important point on the child psychopathology research agenda.

References

Achenbach, T. M. (1991a). *Manual for the Child Behavior Checklist 14–18 and 1991 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1991b). *Manual for the Teacher's Report Form and 1991 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1995). Empirically based assessment and taxonomy: Applications to clinical research. *Psychological Assessment*, *7*, 261–274.

Achenbach, T. M., & Edelbrock, C. (1978). The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin*, *85*, 1275–1301.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213–232.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.; DSM-III). Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., revised; DSM-III-R). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.; DSM-IV). Washington, DC: Author.

Angold, A., Costello, E. J., & Erkanli, A. (1999). Comorbidity. *Journal of Child Psychology and Psychiatry*, *40*, 57–87.

Bannister, D. (1968). The logical requirements of research into schizophrenia. *British Journal of Psychiatry*, *114*, 181–188.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *88*, 588–606.

Bentler, P. M. (1995). *EQS Structural equations program manual*. Encino, CA: Multivariate Software.

Block, J. (1995). A contrarian view of the Five-Factor approach to personality description. *Psychological Bulletin*, *117*, 187–215.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality*. Unpublished Doctoral dissertation, Rijksuniversiteit Groningen, The Netherlands.

Boyle, M. H., Offord, D. R., Racine, Y., Fleming, J. E., Szatmari, P., & Sanford, M. (1993). Evaluation of the Revised Ontario Child Health Study scales. *Journal of Child Psychology and Psychiatry*, *34*, 189–213.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Cantwell, D. P. (1996). Classification of child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry*, *37*, 3–12.

Cantwell, D. P., & Rutter, M. (1994). Classification: Conceptual issues and substantive findings. In M. Rutter, E. Taylor, & L. Hersov (Eds.), *Child and adolescent psychiatry* (3rd ed., pp. 3–21). Oxford: Blackwell Scientific Publications.

Cattell, R. B. (1952). *Factor analysis*. New York: Harper.

Chorpita, B. F., & Barlow, D. H. (1998). The development of anxiety: The role of control in the early environment. *Psychological Bulletin*, *124*, 3–21.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology*, *100*, 316–336.

Clark, L. A., Watson, D., & Reynolds, S. (1995). Diagnosis and classification of psychopathology: Challenges to the current system and future directions. *Annual Review of Psychology*, *46*, 121–153.

Cromwell, R. L. (1975). Assessment of schizophrenia. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology, Vol. 26*. Palo Alto, CA: Annual Reviews.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Eley, T. C., & Stevenson, J. (1999). Using genetic analyses to clarify the distinction between depressive and anxious symptoms in children. *Journal of Abnormal Child Psychology*, *27*, 105–114.

Frick, P. J., Lahey, R. B., Loeber, R., Tannenbaum, L., Van Horn, Y., Christ, M. A. G., Hart, E. A., & Hanson, K. (1993). Oppositional defiant disorder and conduct disorder: A meta-analytic review of factor analyses and cross-validation in a clinic sample. *Clinical Psychology Review*, *13*, 319–340.

Gadow, K. D., & Sprafkin, J. (1994). *Child symptom inventories manual (CSI: parent checklist and CSI: teacher checklist): Screening instruments for childhood emotional and behavioral disorders*. Stony Brook, NY: Checkmate Plus.

Gadow, K. D., & Sprafkin, J. (1997). *Child symptom inventory 4: Norms manual*. Stony Brook, NY: Checkmate Plus.

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research*, *11*, 572–580.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several Five-Factor models. In I. Mervielde, I. J. Deary, F. de Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe, Vol. 7* (pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Hartman, C. A. (2000). *Changing concepts of child psychopathology*. Dissertation, University of Amsterdam, The Netherlands.

Hartman, C. A., Hox, J., Auerbach, J., Erol, N., Fonseca, A. C., Mellenbergh, G. J., Nøvik, T. S., Oosterlaan, J., Roussos, A. C., Shalev, R. S., Zilber, N., & Sergeant, J. A. (1999). Syndrome dimensions of the Child Behavior Checklist and the Teacher Report Form: A critical empirical evaluation. *Journal of Child Psychology and Psychiatry*, *40*, 1095–1116.

Hinden, B. R., Compas, B. E., Howell, D. C., & Achenbach, T. M. (1997). Covariation of the anxious-depressed syndrome during adolescence: Separating fact from artefact. *Journal of Consulting and Clinical Psychology*, *65*, 6–14.

Hox, J. (1998). *Simulcat: Software for simulation of categorical data*. Amsterdam: TT Publicaties.

Hox, J., & Hartman, C. A. (1999). When a poor model meets bad data. In J. Hox & E. D. de Leeuw (Eds.), *Assumptions, robustness, and estimation methods in multivariate modeling* (pp. 141–150). Amsterdam: TT Publicaties.

Hu, L., & Bentler, P. M. (1999). Cut-off criteria for fit indices in covariance structure analysis: Conversational criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Joiner, T. E., Catanzaro, S. J., & Laurent, J. (1996). Tripartite structure of positive and negative affect, depression, and anxiety in child and adolescent psychiatric patients. *Journal of Abnormal Psychology*, *105*, 401–409.

Jöreskog, K. G. (1979). A general approach to confirmatory Maximum Likelihood factor analysis, with addendum. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models* (pp. 21–43). Cambridge: Abt Books.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7 user's reference guide*. Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1993). *New features in LISREL 8*. Chicago: Scientific Software International.

Kamphaus, R. W., & Frick, P. J. (1996). *Clinical assessment of child and adolescent personality and behavior*. Needham Heights, MA: Simon & Schuster.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research*, *25*, 137–155.

Klein, D. F. (1999). Harmful dysfunction, disorder, disease, illness, and evolution. *Journal of Abnormal Psychology*, *108*, 421–429.

Kovacs, M., & Devlin, B. (1998). Internalizing disorders in childhood. *Journal of Child Psychology and Psychiatry*, *39*, 47–63.

Krueger, R. F., Caspi, A., Moffit, T, E., & Silva, P. A. (1998). The structure and stability of common mental disorders (DSM-III-R): A longitudinal-epidemiological study. *Journal of Abnormal Psychology*, *107*, 216–227.

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods*, *4*, 192–211.

Loeber, R., & Keenan, K. (1994). Interaction between conduct disorder and its comorbid conditions: Effects of age and gender. *Clinical Psychology Review*, *14*, 497–523.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.

Macleod, R. J., McNamee, M. A., Boyle, M. H., Offord, D. R., & Friedrich, M. (1999). Identification of childhood psychiatric disorder by informant: Comparisons of clinic and community samples. *Canadian Journal of Psychiatry*, *44*, 144–150.

Meehl, P. (1999). Clarifications about taxometric method. *Applied and Preventive Psychology*, *8*, 165–174.

Newman, D. L., Moffitt, T. E, Caspi, A., Magdol, L., Silva, P. A., & Stanton, W. R. (1996). Psychiatric disorder in a birth cohort of young adults: Prevalence, comorbidity, clinical significance, and new case incidence from age 11 to 21. *Journal of Consulting and Clinical Psychology*, *64*, 552–562.

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin*, *126*, 220–246.

Quay, H. C. (1986a). Classification. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood* (3rd ed., pp. 1–34). New York: Wiley

Quay, H. C. (1986b). A critical analysis of DSM-III as a taxonomy of psychopathoplogy in childhood and adolescence. In T. Millon & G. Klerman (Eds.), *Contemporary directions in psychopathology* (pp. 151–165). New York: Guilford Press.

Rutter, M. (1996). Developmental psychopathology as an organizing research construct. In D. Magnusson (Ed.), *The lifespan development of individuals: Behavioral, neurobiological, and psychosocial perspectives* (pp. 394–413). Cambridge: Cambridge University Press.

Rutter, M., Kebovici, S., Eisenberg, L., Sneznevskij, A. V., Sadoun, R., Brooke, E., & Lin, T. Y. (1969). A triaxial classification of mental disorders in childhood. *Journal of Child Psychology*, *10*, 41–61.

Sandoval, J. (1981). Format effects in two teacher rating scales of hyperactivity. *Journal of Abnormal Child Psychology*, *9*, 202–213.

Seeley, J. R., Lewinsohn, P. M., & Rohde, P. (1997). *Comorbidity between conduct disorder and major depression during adolescence: Impact on phenomenology, associated clinical characteristics, and continuity into young adulthood.* Poster presented at the Eighth Meeting of the International Society for Research in Child and Adolescent Psychopathology, Paris, France.

Shalev, R. S., Hartman, C. A., Stavsky, M., & Sergeant, J. A. (1995). In J. A. Sergeant (Ed.), *Eunethydis: European approaches to hyperkinetic disorder* (pp. 131–147). Zürich, Switzerland: Fotorotar.

Skinner, H. A. (1981). Toward the integration of classification theory and methods. *Journal of Abnormal Psychology*, *90*, 68–87.

Skinner, H. A. (1986). Construct validation approach to psychiatric classification. In T. Millon & G. L. Klerman (Eds.), *Contemporary directions in psychopathology* (pp. 307–330). New York: Guilford Press.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

Swanson, J. H., Sergeant, J. A., Taylor, E., Sonuga-Barke, E. J. S., Jensen, P. S., & Cantwell, D. P. (1998). Attention-deficit hyperactivity disorder and hyperkinetic disorder. *The Lancet*, *351*, 429–433.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, *42*, 233–239.

Tennen, H., Hall, J. A., & Affleck, G. (1995). Depression research methodologies in the Journal of Personality and Social Psychology: A review and critique. *Journal of Personality and Social Psychology*, *68*, 870–884.

Verhulst, F. C., & Van der Ende, J. (1992). Six year stability of parent reported problem behavior in an epidemiological sample. *Journal of Abnormal Child Psychology*, *20*, 595–610.

Vitiello, B., & Stoff, D. M. (1997). Subtypes of aggression and their relevance to child psychiatry. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*, 307–315.

Wakefield, J. C. (1999). Philosophy of science and the progressiveness of the DSM's theory-neutral nosology: Response to Follette and Houts, part 1. *Behaviour Research and Therapy*, *37*, 963–999.

Waldman, I. D., Lilienfield, S. O., & Lahey, B. B. (1995). Toward construct validity in the childhood disruptive behavior disorders: Classification and diagnosis in DSM-IV and beyond. *Advances in Clinical Child Psychology*, *17*, 323–363.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxonomic procedures.* London: Sage Publications.

Wångby, M., Bergman, L. R., & Magnusson, D. (1999). Development of adjustment problems in girls: What syndromes emerge? *Child Development*, *70*, 678–699.

Watson, D., Clark, L. A., Weber, K., Smith Assenheimer, J., Strauss, M. E., & McCormick, R. A. (1995). Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. *Journal of Abnormal Psychology*, *104*, 15–25.

Weiss, B., Suesser, K., & Catron, T. (1998). Common and specific features of childhood psychopathology. *Journal of Abnormal Psychology*, *107*, 118–127.

Zuckerman, M. (1999). *Vulnerability to psychopathology*. Washington, DC: American Psychological Association.

Manuscript accepted 20 February 2001